Journal of Biomedical Informatics 44 (2011) 637–647

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Systems-based biological concordance and predictive reproducibility of gene set discovery methods in cardiovascular disease

Francisco Azuaje [a,*], Huiru Zheng [b], Anyela Camargo [c], Haiying Wang [b]

[a] Laboratory of Cardiovascular Research, Public Research Centre for Health (CRP-Santé), 120, route d'Arlon L-1150, Luxembourg
[b] School of Computing and Mathematics, Computer Science Research Institute, University of Ulster, UK
[c] School of Computing, University of East Anglia, Norwich, NR4 7TJ England, UK

ARTICLE INFO

Article history:
Received 25 August 2010
Available online 17 February 2011

Keywords:
Biomarker discovery
Pathway analysis
Gene set analysis
Cardiovascular diseases
Human heart failure
Disease networks
Translational bioinformatics

ABSTRACT

The discovery of novel disease biomarkers is a crucial challenge for translational bioinformatics. Demonstration of both their classification power and reproducibility across independent datasets are essential requirements to assess their potential clinical relevance. Small datasets and multiplicity of putative biomarker sets may explain lack of predictive reproducibility. Studies based on pathway-driven discovery approaches have suggested that, despite such discrepancies, the resulting putative biomarkers tend to be implicated in common biological processes. Investigations of this problem have been mainly focused on datasets derived from cancer research. We investigated the predictive and functional concordance of five methods for discovering putative biomarkers in four independently-generated datasets from the cardiovascular disease domain. A diversity of biosignatures was identified by the different methods. However, we found strong biological process concordance between them, especially in the case of methods based on gene set analysis. With a few exceptions, we observed lack of classification reproducibility using independent datasets. Partial overlaps between our putative sets of biomarkers and the primary studies exist. Despite the observed limitations, pathway-driven or gene set analysis can predict potentially novel biomarkers and can jointly point to biomedically-relevant underlying molecular mechanisms.

1. Introduction

The discovery of potential biomarkers for detecting, predicting and supporting the treatment of diseases is a central aim to achieve a more personalized, cost-effective medicine [1–4]. The traditional approach to identifying putative biomarkers is based on the premise that genes and proteins act in isolation to contribute to the emergence of specific phenotypes [1,5,6]. This has focused on the detection of differentially expressed genes across clinical conditions (classes) at a pre-defined statistical significance level and after adjusting statistics for multiple-hypothesis testing [7–9].

The notion that common complex diseases rise and progress through the reciprocal actions of multiple genes has motivated the application of advanced methodologies that look at biologically-meaningful "gene sets", rather than lists of independently-assumed genes [10–12]. In this area, different methods based on the definition of scores or indices to summaries the gene expression activity of a gene set at the sample (patient) and class levels have been proposed [13,14]. Such measures encapsulate different aspects of gene set- or pathway-based activity, including statistics based on gene set-class correlations, gene expression means, between-gene correlations, etc. [15,16]. Using this information, researchers can statistically detect differentially expressed gene sets or annotated pathways for supporting biomarker and drug target discovery research [16,17]. The data linked to the genes defining these sets, or the global scores assigned to these sets, have been used as inputs to subsequent discovery tasks [18,19]. Disease classification and prediction is one such task, which aims to develop new computer-based systems for decision support: disease diagnosis, prognosis, prediction of responses to therapies and surrogate end-points in clinical trials [1,20–22].

The possible acceptance of disease biomarkers and derived classification models by the clinical community depends on two crucial factors [23–25]: its prediction or discrimination capacity, and the reproducibility of such performance across patient cohorts or independent datasets. These factors are fundamental initial steps to assess the potential clinical relevance of new biomarkers and

Abbreviations: AUC, area under the receiver operating curve; HF, heart failure; DCM, dilated cardiomyopathy; NF, non-failing hearts; ISC, ischemic heart failure; GEO, gene expression omnibus database; GDS2205, first GEO dataset used in DCM vs. NF classification problem; GDS2206, second GEO dataset used in DCM vs. NF classification problem; GSE1869, first GEO dataset used in ISC vs. NF classification problem; GSE5406, second GEO dataset used in ISC vs. NF classification problem; SVM, support vector machine; GO, Gene Ontology; SS, semantic similarity.

* Corresponding author. Fax: +352 26970 396.
E-mail address: francisco.azuaje@crp-sante.lu (F. Azuaje).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.02.003

their deployment in a hospital setting [26]. The great majority of investigations reported to date have shown great promise with regard to prediction performance [23]. Unfortunately, the task of demonstrating this in independent, multi-site evaluations has proven to be a greater challenge [23]. A reliance on small datasets, experimental sources of error and bias, inconsistencies across laboratory protocols and variance added by data pre-processing procedures have been considered as viable factors influencing poor predictive reproducibility [28–30]. This problem can also be seen as a consequence of the multiplicity of biomarkers or signatures identified by different methods, across different studies dealing with the same clinical problem or even using the same dataset [27,31,32].

Recent research has suggested that the application of gene set analysis may contribute to improving both prediction capacity and reproducibility [33,34]. Moreover, investigations have revealed that, despite the relative small agreement or overlap between biomarkers indentified by different methodologies, gene set analysis can point to significant underlying functional commonalities in the form of shared biological processes [32–35]. Previous studies in this area have focused on datasets originating from cancer research [27,34,35]. Furthermore, some of these studies have concentrated on the characterization of the genes or pathways shared by the disease signatures detected by different methods [34,36]. We set to investigate the problems of predictive performance and reproducibility in the area of cardiovascular research.

We aim to compare different gene set analysis approaches using several independently-generated genome-wide expression datasets from published investigations of human heart failure (HF). We assessed the predictive power of the resulting putative biomarkers through the implementation of classification models, cross-validation and independent evaluations. Furthermore, we look at the biological concordance between methods within and between datasets. Our study shows the robustness of gene set analysis in connection to such integrative, functional relationships. However, a less clear picture unfolds when evaluating classification performance across independent datasets. Notwithstanding these disagreements, we found partial agreements between our investigations and the original (primary) studies using different methodologies. Finally, we show potentially novel biomarkers and associations that will require additional research, and discuss issues possibly influencing the observed discrepancies.

2. Methods

2.1. Research framework

Fig. 1 summarizes the main components and steps of our investigations. After preparing the datasets, we applied five putative biomarker discovery techniques, whose most-statistically significant outcomes (gene sets and gene lists) were analyzed in terms of Gene Ontology (GO) biological process annotations. We searched for statistically significant overlaps between these results within each data and across datasets with the same case-control conditions. Using the top gene sets and genes, we implemented classification models and estimated their performance using leave-one-out cross-validation (LOOCV) on the derivation dataset, and an evaluation on a second, independently-generate dataset. The sequence of this analysis pipeline is illustrated in Fig. 1A. The clinical diagnostic problems, datasets and biomarker discovery methodologies investigated are shown in Fig. 1B, and are described in detail as follows.

2.2. Datasets investigated

We concentrated on two classification problems relating to the diagnosis of human HF. The first problem aims to distinguish between patients with dilated cardiomyopathy (DCM), which is a type of HF, and those with non-failing hearts (NF). The second application deals with the classification of ischemic (ISC) HF vs.

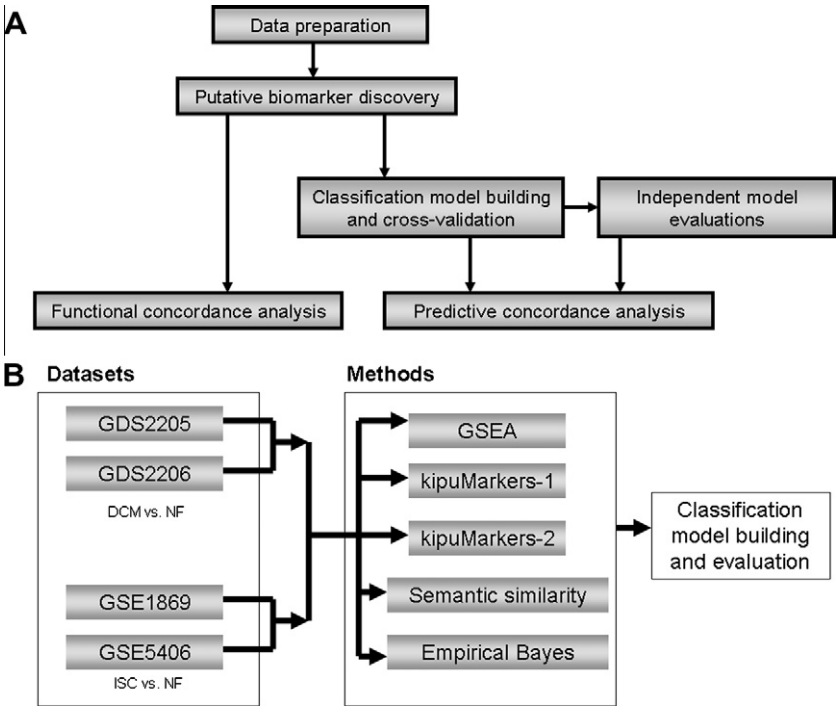

Fig. 1. Overview of research framework. A. Analysis pipeline for assessing functional and predictive concordance of different methods to discover putative biomarkers. B. Description of datasets and biomarker discovery methods investigated. GEO accession numbers are provided for each dataset.

Table 1
Description of datasets investigated. *GEO accession numbers. DCM: dilated cardiomyopathy, NF: non-failing heart, ISC: ischemic heart failure.

Dataset*	Clinical classes	# of genes	# of samples (classes A and B)	Refs.
GDS2205	DCM vs. NF	8021	12 (7, 5)	[36]
GDS2206	DCM vs. NF	21,389	28 (13, 15)	[36]
GSE1869	ISC vs. NF	12,997	16 (10, 6)	[37]
GSE5406	ISC vs. NF	12,997	124 (108, 16)	[38]

NF. In each clinical setting, we analyzed two independently-generated gene expression datasets reported in [37–39]. Normalized data were obtained from the Gene Expression Omnibus database (GEO). We re-scaled expression values across samples so that mean and standard deviation were equal to 0 and 1 respectively. Table 1 lists the main characteristics of these datasets, including their GEO accession numbers and clinical class structure. In total, we analyzed approximately 2.5 million gene expression values across 180 heart tissue samples. In the DCM vs. NF classification problem, the GDS2205 and GDS2206 datasets included 7 and 13 DCM samples respectively, together with 5 and 15 NF samples respectively. In the ISC vs. NF classification application, the GSE1869 (ArrayExpress number: E-GEOD-1869) and GSE5406 (ArrayExpress number: E-GEOD-5406) datasets offered 10 and 108 ISC samples respectively, together with 6 and 16 control (NF) cases. All samples were derived from left ventricle biopsies. From now on, we will refer to these datasets using their accession numbers.

In each clinical application setting, the development of disease classification models consisted of two main phases: *Model derivation* and *independent evaluation*. The derivation phase involves model building (training) and testing through LOOCV. The independent evaluation comprised the application of the models obtained from the derivation phase to a second dataset not included in the derivation phase. In the DCM vs. NF classification problem, GDS2205 and GDS2206 were used as derivation and independent datasets respectively, and then vice versa. The same dataset selection–evaluation procedure was implemented in the ISC vs. NF biomarker discovery application.

2.3. Putative biomarker discovery methods investigated

Five published techniques for the identification of putative biomarker discovery using gene expression data were investigated (Fig. 1B): Gene Set Enrichment Analysis (GSEA) [15], two pathway-centric analysis techniques from the *kipuMarkers* methodology [16], *SimTrek*: a GO-based similarity gene set analysis technique [40], and empirical Bayes analysis for detecting differentially expressed genes [41].

For a given gene set, G: $\{g_1, g_2, \ldots, g_n\}$, the GSEA method estimates a enrichment score for G based on the Kolmogorov–Smirnov statistic [15], which reflects the differential correlation between G and the phenotype classes under consideration. GSEA is one of the most-used gene set analysis techniques currently available. False discovery rates for each G detected are estimated with a phenotype-based permutation procedure, as detailed in [15]. We concentrated our functional and classification model building analyses on the five most significant gene sets (FDR $\leqslant$ 0.01).

Given G, the *kipuMarkers-1* technique maps sample-specific gene expression data on the input pathways, and scores each pathway with the mean expression values observed in each sample. The pathway-based scores, also referred to as *gene set activity levels*, are then used to detect differential associations with the phenotypes studied. Differential pathway-centric expression levels are quantified as *perturbation scores*, and are computed using the t-statistic. Statistical control for multiple-testing is implemented using a permutation procedure [16]. A variation of this methodology, *kipuMarkers-2*, estimates a pathway activity level by computing the mean of the pair-wise differences of expression values observed in G, normalized to the size of the pathway. Perturbation scores for each pathway and statistical testing are implemented as in *kipuMarkers-1*. We concentrated our functional and classification model building analyses on the five most significant gene sets (FDR $\leqslant$ 0.01).

In this research GSEA, kipuMarkers-1 and kipuMarkers-2 used a collection of 639 molecular pathways from the Molecular Signatures Database (MSigDB) [15]. Our predictions concentrated on its "C2 collection" of canonical pathways, which integrates annotated signaling and metabolic pathways obtained from different manually-curated databases.

Unlike GSEA and kipuMarkers, the SimTrek technique does not require gene sets as user-defined inputs. Given a seed list of gene products, G: $\{g_1, g_2, \ldots, g_n\}$, gene sets: $G_{g1}, G_{g2}, \ldots, G_{gn}$, are assembled for each gene in G using the GO-based functional similarity between g_i and other genes in the genome. The nearest neighbors in this "semantic" space are retrieved as the predicted gene set for each query gene. Thus, for a given list of query genes, SimTrek predicts gene sets that can be subsequently analyzed using expression data corresponding to their constituent genes. SimTrek computes whole-genome, GO-based functional similarity using information theoretic methods, as explained in [40]. SimTrek has shown its potential to infer relevant gene sets, including putative protein–protein interactions [40]. In this study, SimTrek processed the 10 genes most differentially expressed as its input queries. Differentially expressed genes were identified with the empirical Bayes method [41]. For each query gene, SimTrek generated gene sets consisting of the 10-nearest neighbors in the GO-based similarity space, using non-IEA (not Inferred from Electronic Annotations) Biological Process GO terms and human genes only. Thus, SimTrek predicted for each dataset (up to) 10 gene sets, each comprising (up to) 10 putative biomarkers.

To expand our investigation beyond gene set analysis, we report differentially expressed genes as putative biomarkers identified by a linear model method based on the Empirical Bayes moderate t-statistic [41]. After fitting a linear model to the expression profile of each gene in a dataset, standard errors are moderated applying an Empirical Bayes model. A moderated t-statistic and a log-odds of differential expression is then computed for each contrast for each gene. Empirical Bayes has shown predictive robustness even in small size datasets [41], and has been successfully applied to support systems biology research [42]. Here we focused on the top-100 differentially expressed genes detected by this method, with all of them reporting $P < 0.01$.

2.4. Bioinformatic tools and statistics

Analyses with GSEA, kipuMarkers and SimTrek were implemented with Java-based software provided by their authors [15,16,40]. Linear models based on moderated t-statistics and Empirical Bayes analyses were implemented using the R software package, Limma library from Bioconductor [43]. Statistically detectable overlaps between putative biomarker sets, in terms of their GO annotations, were estimated with the Java-based software ToppCluster [44]. This system applies the hypergeometric test and multiple-testing corrections to detect statistically detectable GO term enrichments. We focused on significant enrichments with FDR $\leqslant$ 0.01. All other statistical measures reported were adjusted for multiple-testing as indicate above. Classification models were implemented with the Weka data mining platform [45,46]. Due to its demonstrated classification performance and robustness, we focused on models based on linear Support Vector Machines

(SVM), trained with Platt's sequential minimal optimization algorithm (complexity parameter = 1, exponent = 1) [46,47]. Classification performance was summarized with areas under the receiving operating characteristic curves (AUCs). Associations between sets of putative biomarkers and prior biomedical knowledge were examined using published literature and Pubmed.

3. Results

3.1. Discovery of clinically-relevant biological pathways

In GDS2205, the gene set analysis methods detected diverse significantly altered pathways, as annotated in the MSigDB: From metabolism, immune responses to protein synthesis (Table S1 in Supplementary information). Overall GSEA, kipuMarkers-1 and kipuMarkers-2 shared metabolism-related pathways as the most significant perturbations. Unlike GSEA, the kipuMarkers techniques positioned immune response pathways (such as those involved in immune cell control) at the top of the most differentially altered gene sets across DCM and NF patients. GSEA top-ranked pathways more specifically involved in protein biosynthesis and energy metabolism (Table S1). In GDS2206, this preference toward metabolic, immune response and protein synthesis pathways was preserved (Table S2). Nevertheless, kipuMarkers-1 and kipuMarkers-2 highlighted the alteration of cell death processes (e.g. apoptosis and CD40 pathways). GSEA ranked specific pathways involved in amino acid metabolism as top perturbations underlying HF (e.g. degradation of valine and leucine).

In GSE1869, kipuMarkers-1 detected strong perturbations in calcium regulation in cardiac cells and smooth muscle contraction (as defined in the MSigDB) (Table S3). kipuMarkers-v2 highlighted differential alternations in molecular signaling processes known to be implicated in cancers (e.g., estrogen signaling and cell migration of carcinoma cells, Table S3). GSEA emphasized the deregulation of pathways responsible for transmembrane signal transduction (e.g., G protein-coupled receptor interactions). kipuMarkers-1 detected major alterations in pathways relevant to sugar metabolism (e.g., streptomycin biosynthesis and starch metabolism). kipuMarkers-v2 and GSEA pointed to pathways implicated in inflammation and responses to infections (e.g., CCR3 and cytokine interaction pathways). In GSE5406, all the methods indentified significant alternations in amino acid synthesis and energy metabolism pathways (Table S4). In the list of top perturbations, these methods also positioned pathways directly implicated in cell division and migration, including those deregulated in different cancers, e.g., ERK and RECK. These methods also detected important perturbations in pathways required for cellular death (e.g., Tall 1 pathway). GSEA was capable to pinpoint heart-specific perturbations, e.g. insulin receptor and PIP3 signaling in cardiac myocytes.

3.2. Intra-dataset, process-oriented concordance of putative biomarker discovery methods

In each dataset we performed a closer examination of potentially significant functional overlaps between the methods on the basis of shared GO Biological Process terms. We look at the statistical enrichment of GO terms observed in the top-perturbed pathways reported above. This analysis also included the putative biomarkers detected by the methods not driven by pathway analysis (semantic similarity and empirical Bayes). Figs. 2 and 3 illustrate examples of such a functional concordance within the same datasets. In these figures, methods and GO Biological Process terms are depicted with circles and squares, respectively. Lines are used to show significant statistical associations between methods and the (top-10) shared GO terms (FDR $\leqslant$ 0.01). The darker the line the stronger the GO term enrichment detected by a method. Alternative visualizations of these figures are included in the Supplementary information (Tables S8 and S9).

In both GDS2205 and GDS2206 datasets (DCM vs. NF classification), stronger agreements between kipuMarkers-1, kipuMarkers-2 and GSEA can be observed (Fig. 2). Semantic similarity and empirical Bayes showed relatively little overlaps between them and with the pathway-driven methods. These results corroborated, for example, the shared capacity of kipuMarkers-1, kipuMarkers-2 and GSEA to point to the significant deregulation of metabolic and immune processes.

Similar functional overlapping patterns were observed in GSE1869 and GSE5406 (ISC vs. NF classification, Fig. 3). kipuMarkers-1, kipuMarkers-2 and GSEA tended to point to similar perturbations. In these cases, however, the semantic similarity method seemed to have more commonalities with kipuMarkers-1, kipuMarkers-2 and GSEA in comparison to the previous analyses, and to the empirical Bayes method. The strongest agreements between all the five methods were observed in processes required for energy metabolism, immune responses and signal transduction (GSE1869, Fig. 3A). In GSE5406 (Fig. 3B), a strong convergence of pathway-driven methods was observed in processes related to cell proliferation and cell death. In the list of most significantly enriched GO terms, semantic similarity and empirical Bayes techniques showed relatively weak concordance between them and in relation to the other methods.

3.3. Inter-dataset, process-oriented concordance of putative biomarker discovery methods

After investigating the intra-dataset agreement between methods, we characterized each dataset according to the GO Biological Process term enrichments in the different gene lists and gene sets detected by the methods. This allowed us to estimate statistically detectable overlaps between datasets independently of the method applied. Despite the multiplicity of genes and gene sets detected by each method within each dataset, we found strong functional concordance between them, within each disease classification problem. Fig. 4 and Table S10 (Supplementary information) illustrate these relationships for the 10-most statistically over-represented GO terms.

In the DCM vs. NF classification problem, different statistically detectable functional overlaps between the putative biomarkers from GDS2205 and GDS2206 were found (Fig. 4A). Examples of these commonalities include significant alterations in processes implicated in cell death, metabolism and phosphorylation (FDR < 0.01). In the ISC vs. NF classification problem (Fig. 4B), GSE1869 and GSE5406 allowed the extraction of putative biomarkers that are strongly implicated in common biological processes: immune responses, cell proliferation, cell death and development (FDR < 0.01).

3.4. Disease classification capacity of putative biomarker discovery methods

We tested the disease classification capacity of the genes and gene sets detected as significantly, highly differentially expressed. We built and tested support vector machine models based on the expression data from the genes defining an individual pathway (GSEA) and gene set activity levels (kipuMarkers-1 and kipuMarkers-2). We also evaluated classifiers based on different combinations of genes detected by semantic similarity and empirical Bayes methods. Models were built and tested (with LOOCV) on each dataset separately (Section 2). Classification performances above an AUC = 0.80 (95% confidence interval, CI: 0.54–1.0 (GDS2205), 0.63–0.97 (GDS2206), 0.56–1.0 (GSE1869), and

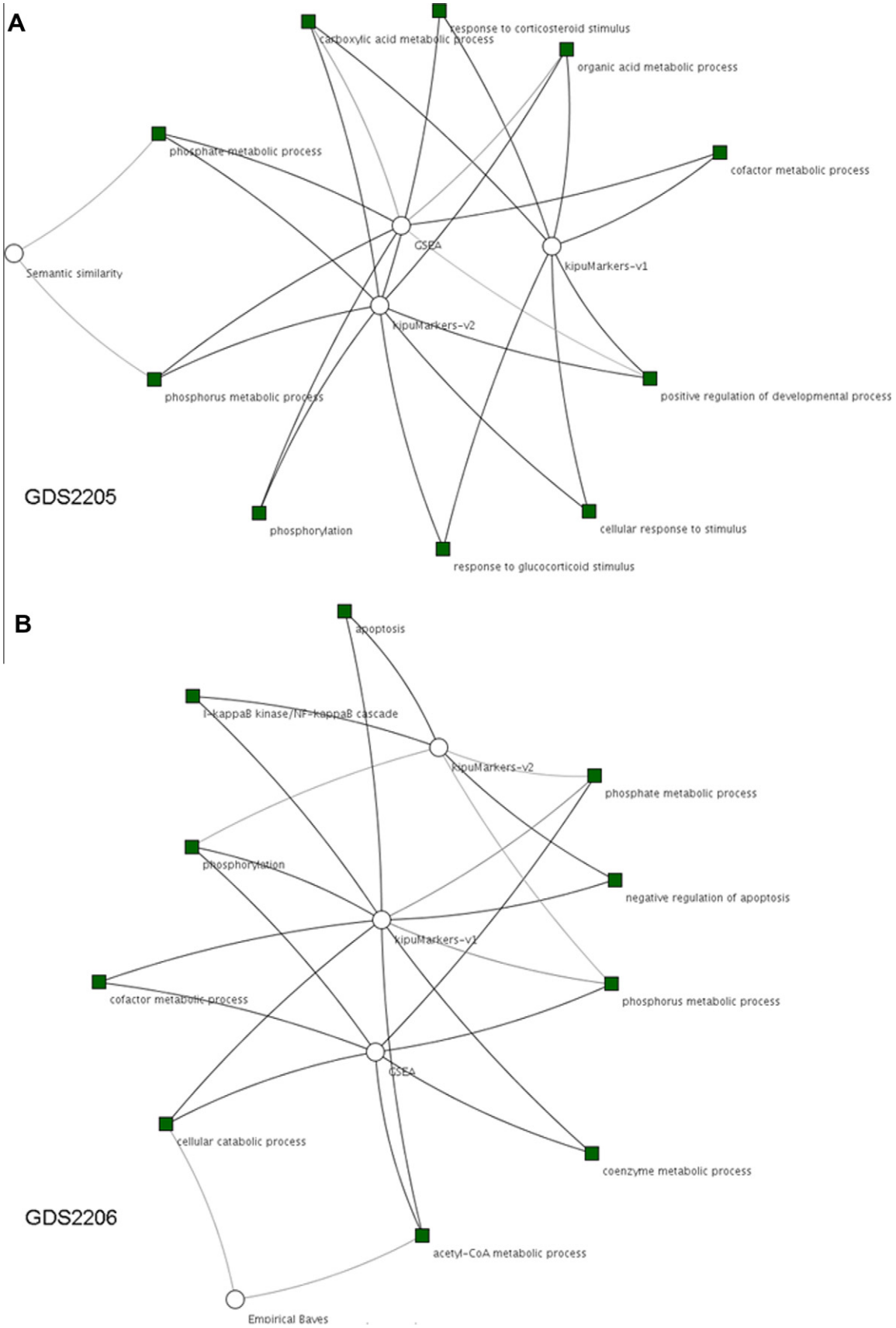

Fig. 2. Functional concordance of putative biomarker discovery methods within the same (DCM vs. NF) dataset. A. GDS2205 dataset. B. GDS2206 dataset. Methods and GO Biological Process terms are depicted with circles and squares respectively. Lines show significant statistical associations between methods and the (top-10) shared GO terms (FDR $\leqslant$ 0.01). The darker the line the stronger the GO term enrichment detected.

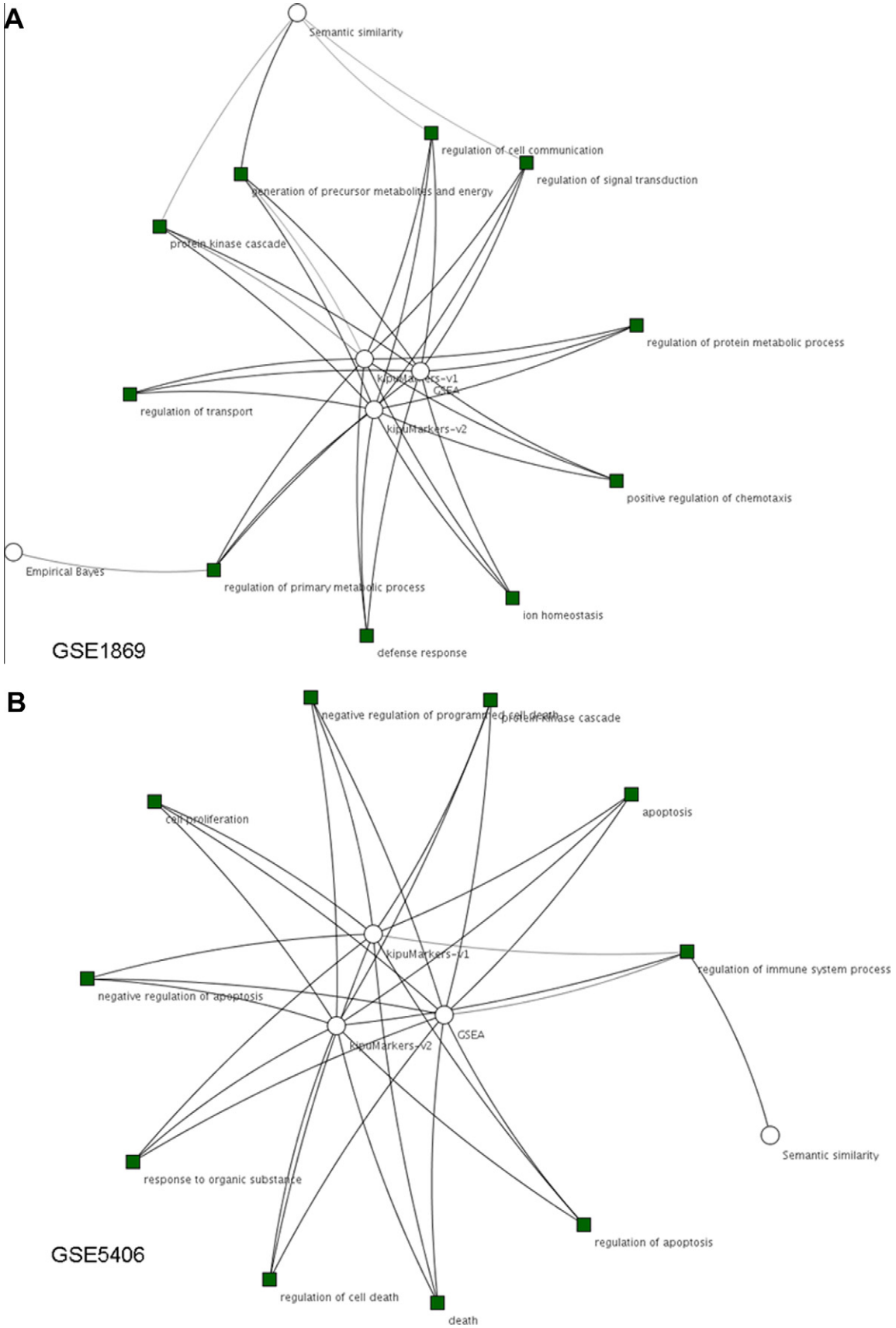

Fig. 3. Functional concordance of putative biomarker discovery methods within the same ISC vs. NF dataset. A. GSE1869 dataset. B. GSE5406 dataset. Methods and GO Biological Process terms are depicted with circles and squares respectively. Lines show significant statistical associations between methods and the (top-10) shared GO terms (FDR $\leqslant$ 0.01). The darker the line the stronger the GO term enrichment detected.

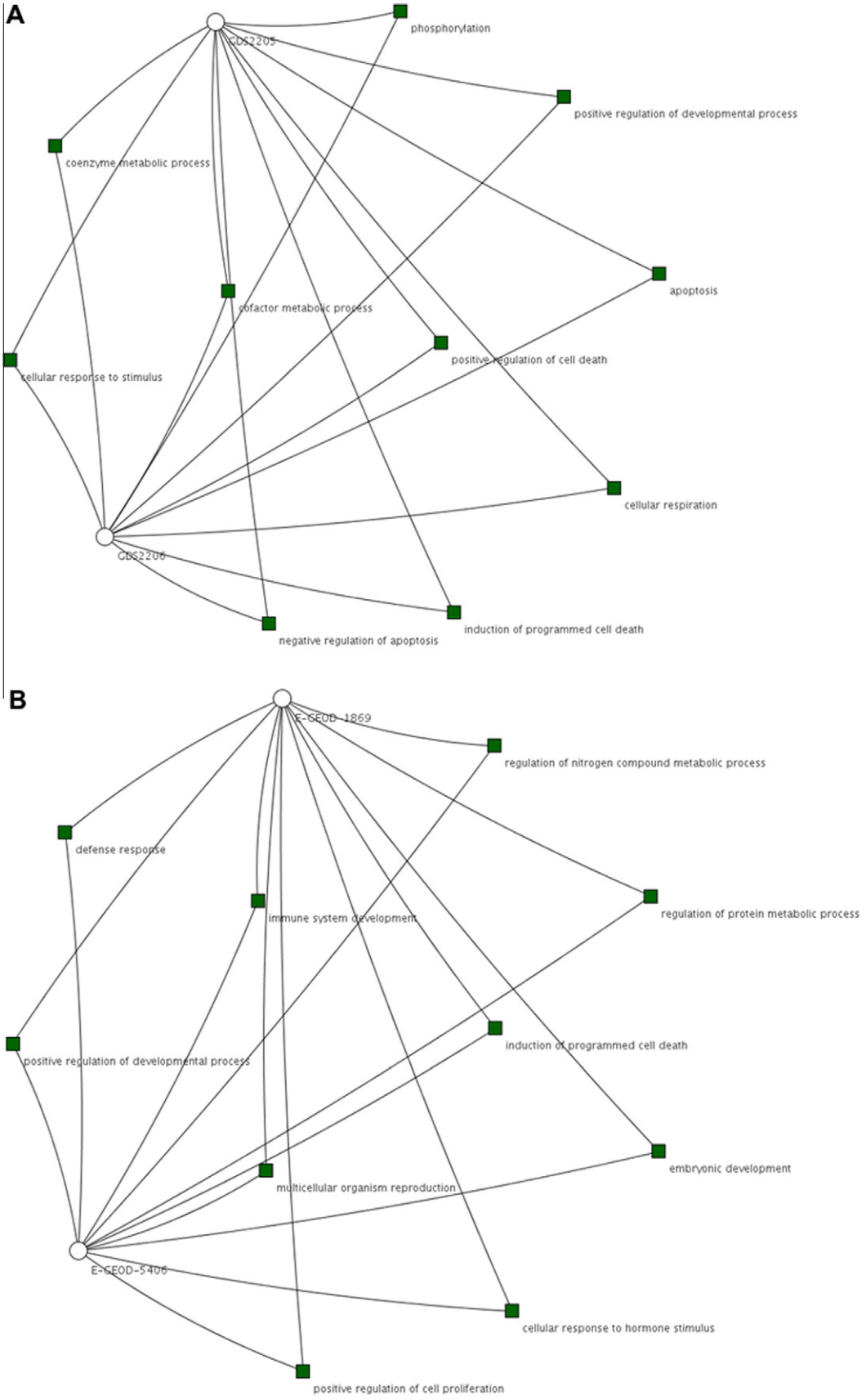

Fig. 4. Functional concordance of putative biomarker discovery methods across datasets. A. DCM vs. NF datasets. B. ISC vs. NF datasets. Datasets and GO Biological Process terms are depicted with circles and squares respectively. Lines show significant statistical associations between datasets and the (top-10) shared GO terms (FDR $\leqslant$ 0.01). The darker the line the stronger the GO term enrichment detected.

0.66–0.93 (GSE5406)) were obtained for all methods and datasets. Perfect discrimination capacity (AUC = 1) was achieved by the models based on putative biomarkers detected by: kipuMarkers-2 technique (GDS2205, GDS2206, GSE1869 datasets) using its top-5 pathway activity levels as inputs; kipuMarkers-1 (GSE1869) using its top-5 pathway activity levels as inputs, GSEA (GDS2205) using gene expression data from the genes defining its top-5 pathways, and semantic similarity (GDS2205) using gene expression data from the top-5 differentially expressed genes and their corresponding top-5 nearest neighbors (Section 2). Perfect classification performance was also observed from relatively simpler models based on combinations of genes detected by empirical Bayes: in GDS2205 (inputs: TRMT5, C14orf133, C16orf45, PPP2R4), GDS2206 (inputs: COPS8, RAB28, SKP1, ACAD10, C5orf23), and GSE1869 (inputs: STAT6, FCN3, C22orf9, PHLDA1, ENDOGL1, DEFB126, CCDC93). These representative classification performances are graphically illustrated in Fig. 5 (first column of heat maps) and described in more detail in Table S5.

3.5. Cross-dataset prediction reproducibility: Independent evaluations of methods

We assessed the predictive reproducibility of the models reported above through their evaluation on independent datasets, as specified in Section 2. Fig. 5 summarizes maximum, representative results for all putative biomarker discovery methods and datasets investigated. The first columns of the heat maps depict the classification performance of the models on the derivation dataset (Section 3.4), while the second columns illustrate the classification performance (AUC) of the resulting models when tested on the independent dataset. Table S6 shows details of these independent evaluations. Overall, poor reproducibility of classification performance was observed. In some cases, classification performance equal or below random classification was observed. The maximum classification performance observed in these independent validations was obtained with the genes retrieved by the semantic similarity method (AUC = 0.92, 95% CI: 0.75–1.0), using GSE5406 and GSE1869 as derivation and independent datasets respectively. This was followed by the model built with the top-5 pathway activity indices extracted by the kipuMarkers-v1 technique (independent evaluation: AUC = 0.90, 95% CI: 0.80–1.0), with GSE1869 and GSE5406 as derivation and independent datasets respectively.

In the DCM vs. NF classification problem (Fig. 5A), the empirical Bayes and semantic similarity techniques reported the most consistent classification performances between derivation and independent evaluation results. In the case of the empirical Bayes method, maximum classification performances with AUC = 0.83 (95% CI: 0.67–0.99) and 0.75 (95% CI: 0.5–1.0) were estimated when GDS2206 and GDS2205 were used as independent datasets respectively. The most robust prediction performance of the models based on the biomarkers detected by the semantic similarity method was observed when using GDS2205 and GDS2206 as derivation and independent datasets respectively (independent evaluation: AUC = 0.87, 95% CI: 0.73–1.0).

In the ISC vs. NF classification problem (Fig. 5B), the empirical Bayes and semantic similarity techniques again provided the inputs to the most reproducible classification models. Similar classification performances were obtained for the model derivation (LOOCV) and independent evaluations when GSE5406 was used as the derivation dataset. The models based on inputs from the empirical Bayes and semantic similarity methods reported AUC = 0.75 (95% CI: 0.5–1.0) and 0.92 (95% CI: 0.76–1.0) respectively when tested on the independent dataset. In contrast to the DCM vs. NF classification application, an improvement in classification reproducibility was observed in the case of our pathway-driven analysis methods. GSEA and kipuMarkers-1 showed relatively good reproducibility when using GSE1869 as derivation dataset (AUC = 0.79, 95% CI: 0.65–0.93, and AUC = 0.90, 95% CI: 0.80–1.0, respectively on independent dataset).

Although the focus of this study was to assess predictive concordance of gene set-based models, we also examined performances obtained from single genes detected by empirical Bayes to expand our comparisons. The vast majority of independent evaluations using single-genes reported very low classification performance. Indeed, most of the top single-genes derived from all datasets provided classification performance equal or close to random classification (AUC = 0.5) when tested on the independent datasets. Models independently evaluated on GDS2206 and GSE1869 illustrated exceptions to this observation. For example, in the former scenario, those built on single top-genes, such as TRMT5 and HMGN2, reported AUC $\geqslant$ 0.70. Fig. S1 (Supplementary information) depicts the ROC curves from models based on these genes.

3.6. Predictive agreement with primary investigations

We examined the papers that originally reported the datasets investigated here to assess the predictive concordance with our

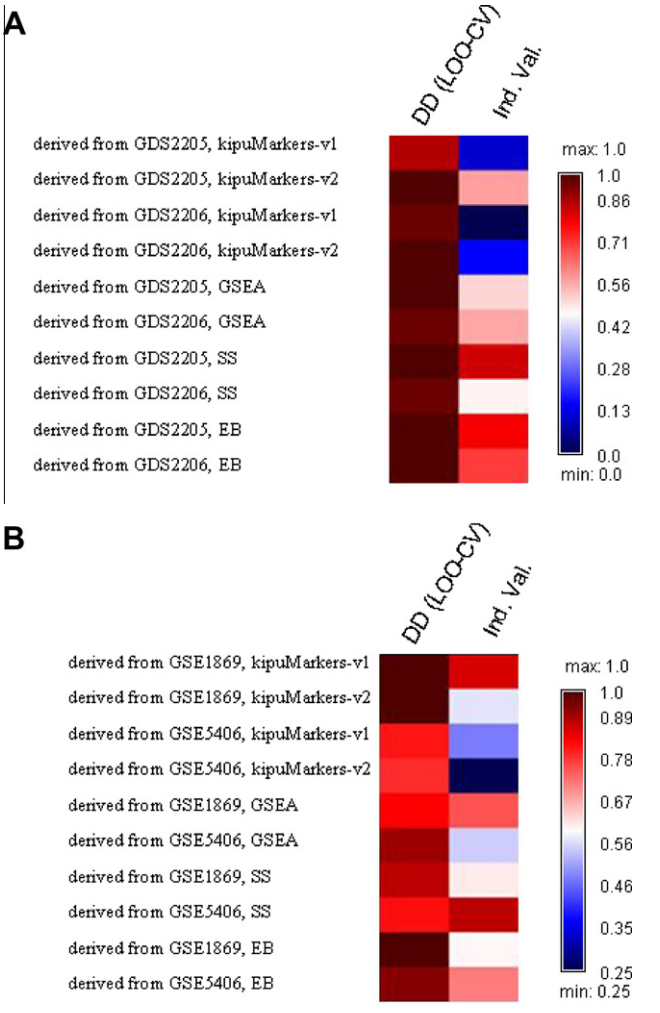

Fig. 5. Visualization of maximum classification performance across different datasets, model input discovery techniques and model derivation–evaluation settings. A. DCM vs. NF datasets. B. ISC vs. NF datasets. Heat maps are used to illustrate the classification performance estimated with AUC values. First map column depicts results from model derivation datasets (with LOOCV). Second column shows the results from independent evaluations using the complementary dataset (not used for model derivation). All models were based on linear SVMs. SS: semantic similarity method, EB: empirical Bayes.

results, independently of the putative biomarker technique applied. The largest agreement in terms of overlapping genes identified as putative biomarkers was observed between our results and the research on GSE5406 [39]. The primary research applied Significance Analysis of Microarray (SAM) to detect potential biomarkers resulting in 642 differentially expressed genes. We then compared their signature with our most differentially expressed pathways detected by kipuMarkers, GSEA and semantic similarity techniques, as well as the differentially expressed genes detected by empirical Bayes. A total of 102 genes were found in common between our investigations: 13 genes detected by kipuMarkers, 18 genes included in the pathways detected by GSEA, six from empirical Bayes and seven detected by the semantic similarity method. We could not compare classification performances as the primary study did not implement this task. Similar comparisons were completed using the other datasets as follows. Table S7 offers details of these relationships.

In terms of the putative biomarkers identified, disagreements between our analyses and the primary research were observed. However, there is indication of classification performance concordance. In the primary study [37], GDS2205 was analyzed with SAM and 27 significantly differentially expressed genes were detected. Only the GSEA and the kipuMarkers methods reported genes in common (one gene each from the most significant pathway) with this biosignature. Barth et al. [37] implemented classification models with the Prediction Analysis for Microarrays (PAM) technique, and obtained an average classification accuracy of 90%, which is similar to the performances observed above. GDS2206 showed similar relationships with the original research: 27 differentially expressed genes (SAM), maximum classification accuracy of 90%. In this case the top pathway detected by kipuMarkers and the set of most differentially expressed genes found by empirical Bayes included only one gene (each) in common with the original report. The primary investigation of GSE1869 [38] reported 31 genes significantly differentially expressed using SAM. Classification performance analysis of these potential biomarkers was not conducted, and their overlap with our top predictions was small: 1 (GSEA), 2 (kipuMarkers), 1 (empirical Bayes) and 1 (semantic similarity).

4. Discussion and conclusions

We investigated the predictive and functional concordance of five techniques for the discovery of putative biomarkers, across independent datasets derived from cardiovascular research studies. We found little commonality in terms of the genes and gene sets (pathways) defining the most significant, potential biosignatures. This multiplicity of putative biomarkers was also observed when compared to the original studies reporting the datasets investigated. Such lack of overlaps has been reported using data derived from cancer research [27]. We also found that the classification performances obtained from this diversity of potential biosignatures tend to be high when estimated on each dataset. Such performances were, in general, difficult to replicate in independent datasets. However, comparable prediction performances between derivation and independent validation analyses were observed in some settings, with independent evaluations reporting AUC > 0.75. This suggests, on one side, the feasibility of successfully validating and reproducing prediction results across independent datasets. On the other side, our results suggest that some of our models, and previous research, over-fitted derivation datasets. Despite this lack of predictive consistency, we also found that putative biomarker identification methods can detect gene sets significantly implicated in common, specific biological processes within and across datasets. Moreover, independent studies (datasets) can be linked through such functional concordance.

Our results showed that classification performance reproducibility may be influenced by both the observed multiplicity of potential biomarkers and the sizes of the derivation datasets. However, we do not have sufficient evidence to suggest that relative small sizes of data can act as the major cause of lack of reproducibility. In this study, we found that relatively small datasets could actually offer robust predictions across independent datasets (e.g. GDS2205). Furthermore, lack of reproducibility can also be observed when using larger model derivation datasets (GSE5406, 124 samples). This adds to the notion that lack of predictive reproducibility and diversity of putative biomarkers is governed by a variety of experimental and biological factors. A significant amount of this variability may be explained by the biological redundancy and complexity inherent in the molecular networks underpinning common diseases [10]. Potential sources of heterogeneity also involve the bias and variability added by the classification models depending of dataset sizes and characteristics, as well as the presence of unknown confounding factors including those related to biomedical history and sub-phenotypes [30]. Moreover, there are several experimental factors that go back to the data generation and preprocessing steps: correlated noise between genes on a microarray experiment, errors and noise introduced during the sample preparation and RNA amplification phase, and the bias or inconsistencies contributed by different normalization or pre-processing techniques [29,48]. Also recent research has demonstrated that model performance and its reproducibility can depend on the clinical endpoint or class under investigation [32], and that prediction capability may depend on the combined effect of classification complexity and sample size [49]. These factors may be used to explain the relative lack of predictive reproducibility and deserve future investigations.

In GSE5406, we obtained poor classification generalization. However, we also observed strong overlaps, in terms of putative biomarkers selected, with the original study that generated this dataset. Thus, this application may represent an example of a classification problem in which data size may not be a dominant factor to achieve predictive robustness. This is in line with evidence, including that recently presented by the latest assessment of the MAQC Consortium [32], that indicates that data size alone may not always matter. This is supported by the fact that it is possible to obtain robust models using relatively small datasets, and poor generalization when using larger datasets. This corroborates the understanding that predictive reproducibility is a multi-factorial, complex problem in which multiple technical and domain-specific obstacles need to be overcome. For example, recent research based on a wide variety of dataset sizes also showed that data size should not be seen as a dominant factor in isolation. In particular, its interplay with biological classification problem complexity and selection of prediction end-points has been shown [49].

4.1. Potential biomedical novelty of our results

To estimate the potential biomedical novelty or relevance of our predictions, we searched the literature and databases for known gene-disease relationships. Examples of relevant associations detected by the empirical Bayes method include: STAT6 (GSE1869 dataset), which has been linked to myocardial ischemia and reperfusion [50], and HMOX2 (GSE5406 dataset) with no established connection to HF, but which is known to be induced by oxidative stress and inflammation [51]. GSEA detected the SIG_PIP3 pathway, which is a 63-gene signaling pathway implicated in cardiac myocytes, as significantly differentially regulated in GSE5406 (FDR = 0.001). In GDS2205, the semantic similarity method identified a number of deregulated genes known to be involved in heart

physiology. For example, HSF2 is required for mouse heart development [52], and KLF5 has been associated with cardiac hypertrophy and fibrosis [53]. In GDS2205, kipuMarkers-v1 indentified the acetaminophen pathway as the most differentially deregulated pathway (Table S1). Previous research in animal models suggested acetaminophen as a safe drug in the context of post-myocardial infarction, though no major cardioprotective properties were observed [54]. In GSE5406, the Reck pathway was identified by kipuMarkers-2 as a top putative target (Table S1). This pathway is responsible for membrane-anchored inhibition of matrix metalloproteinases, including MMP9 and TIMP1, which have been associated with heart failure [55].

4.2. Limitations

The size and depth of predictions examined may be seen as a possible limitation of our study. We reported findings that focused on the most significant gene sets and pathways detected. This means that potentially relevant groups of genes could have been ignored. However, by concentrating on our top predictions we also aimed to reduce the potential number of false positive associations. Moreover, we were interested in investigating biosignatures that were identified as the most biologically-promising or statistically-supported by the methods evaluated.

Also we are naturally limited by the quality and coverage of the molecular pathways used in our analysis. Nevertheless, we aimed to discover putative pathway-based biomarkers using annotated, experimentally-validated gene sets. Moreover, we also incorporated a discovery method fully driven by linear models of available data, which does not rely on pathways selected a priori. In any case there is a need to implement comparisons between network-based methods that are not based on pre-defined gene sets, such as those reported in [12,18,19,58,59]. Future studies may also include alternative gene set discovery techniques, such as those reviewed in [60].

Another constraint may be represented by our emphasis on classification models based on SVM. However, we did not intend to perform a comprehensive comparison of classification techniques in gene expression data, which has been reported elsewhere [56,57], and therefore decided to focus on a known powerful and robust approach [56]. Future research could provide deeper views of predictive concordance with an emphasis on classification model diversity. We also acknowledge that future investigations in the area of cardiovascular diseases should include other datasets and clinical conditions aside from heart failure.

Gene set selection for classification was performed on the derivation datasets, and the classification performance of the resulting sets was then tested on the independent evaluation datasets, which were not used in the prior phase. The LOOCV was used to provide estimates of classification performance of selected biomarkers within the derivation dataset only. Also note that the selection of biomarkers within the derivation datasets was not wrapped around the classifiers implemented. Regardless of these bias prevention strategies, we acknowledge that the use of LOOCV may have contributed to a biased estimate of the classification performance assigned to models within the derivation datasets only. However, also note that in some of the classification applications we obtained relatively consistent classification performances across derivation-independent datasets, which suggest that at least in some cases the estimates were not overoptimistic.

Another critical research topic is to assess how data preprocessing can influence gene set or network-based biomarker discovery. In this research, we used data already normalized by the original studies. This allowed us to make comparisons with the original research's findings. Future studies will require the use of raw data files to estimate pre-processing effects. Moreover, the importance of applying a common normalization procedure across dataset merits consideration [61].

5. Conclusions

We offered evidence that a diversity of putative biomarker discovery methods and derived sets of biosignatures detected in independent datasets can point to common molecular mechanisms, which may characterize or control clinical phenotypes. This was demonstrated within and between independent datasets. However, a less clear picture is revealed with regard to classification performance reproducibility across and within datasets. In this case, discrepancies and lack of predictive generalization was a common theme, though some methods and datasets showed to be the exception. Partial overlaps and consistency were also found in relation to the studies originally reporting the datasets.

To the best of our knowledge, this is the first functional and predictive concordance analysis performed using independent datasets in the area of human heart failure. In particular, we addressed the problem of pathway-based functional concordance and of classification performance reproducibility in the specific context of gene set discovery. Despite the limitations and constraints of gene set analysis techniques, they represent useful tools for suggesting putative biomarkers and for supporting the elucidation of key mechanisms with promising causal or correlative implications. The successful independent validation of disease biomarker models may be benefited by considering multiple discovery techniques, and by ensuring that common data acquisition and pre-processing standards are applied across research sites. Furthermore, new strategies will be required to manage the intrinsic, biological variability encoded in the underlying mechanisms of complex common diseases. This may require moving from the idea of detecting differentially expressed genes or pathways, to strategies specifically aiming to infer genes with potential causal roles, such as those acting as master regulators.

To sum up our findings, different "gene set" methods for discovering putative biomarkers can provide concordant predictions in the sense that they can point to common critical molecular pathways underlying the disease investigated. However, in terms of classification performance this consistency is less clear. This suggests that additional research is needed about approaches to incorporate "gene set" analysis for disease classification, such as the measurement of integrated pathway expression activity.

6. Conflict of interest

None declared.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jbi.2011.02.003.

References

[1] Azuaje F, Devaux Y, Wagner D. Computational biology for cardiovascular biomarker discovery. Brief Bioinform 2009;10:367–77.
[2] Giljohann DA, Mirkin CA. Drivers of biodiagnostic development. Nature 2009;462:461–4.
[3] Lussier YA, Butte AJ, Hunter L. Current methodologies for translational bioinformatics. J Biomed Inform 2010;43:355–7.
[4] Butte AJ. Translational bioinformatics applications in genome medicine. Genome Med 2009;1:64.
[5] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23:2507–17.
[6] Sinha A, Hripcsak G, Markatou M. Large datasets in biomedicine: a discussion of salient analytic issues. J Am Med Inform Assoc 2009;16:759–67.

[7] Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA 2003;100:9440–5.
[8] Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003;19: 368–75.
[9] Chen R, Sigdel TK, Li L, Kambham N, Dudley JT, Hsieh SC, et al. Differentially expressed RNA from public microarray data identifies serum protein biomarkers for cross-organ transplant rejection and other conditions. PLoS Comput Biol 2010;6:pii:e1000940.
[10] Schadt EE. Molecular networks as sensors and drivers of common human diseases. Nature 2009;461:218–23.
[11] Choi Y, Kendziorski C. Statistical methods for gene set co-expression analysis. Bioinformatics 2009;25:2780–6.
[12] Chuang HY, Lee E, Liu YT, Lee D, Ideker T. Network-based classification of breast cancer metastasis. Mol Syst Biol 2007;3:140.
[13] Hung JH, Whitfield TW, Yang TH, Hu Z, Weng Z, DeLisi C. Identification of functional modules that correlate with phenotypic difference: the influence of network topology. Genome Biol 2010;11:R23.
[14] Irizarry RA, Wang C, Zhou Y, Speed TP. Gene set enrichment analysis made simple. Stat Methods Med Res 2009;18:565–75.
[15] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005;102:15545–50.
[16] Azuaje F, Devaux Y, Wagner DR. Integrative pathway-centric modeling of ventricular dysfunction after myocardial infarction. PLoS ONE 2010;5:e966.
[17] Mani KM, Lefebvre C, Wang K, Lim WK, Basso K, Dalla-Favera R, et al. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. Mol Syst Biol 2008;4:169.
[18] Lefebvre C, Rajbhandari P, Alvarez MJ, Bandaru P, Lim WK, Sato M, et al. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. Mol Syst Biol 2010;6:377.
[19] Carro MS, Lim WK, Alvarez MJ, Bollo RJ, Zhao X, Snyder EY, et al. The transcriptional network for mesenchymal transformation of brain tumours. Nature 2010;463:318–25.
[20] Azuaje F. What does systems biology mean for biomarker discovery? Expert Opin Med Diagn 2010;4:1–10.
[21] Deo RC, Hunter L, Lewis GD, Pare G, Vasan RS, Chasman D, et al. Interpreting metabolomic profiles using unbiased pathway models. PLoS Comput Biol 2010;6:e1000692.
[22] Plump AS, Lum PY. Genomics and cardiovascular drug development. J Am Coll Cardiol 2009;53:1089–100.
[23] Koscielny S. Why most gene expression signatures of tumors have not been useful in the clinic. Sci Transl Med 2010;2:14ps2.
[24] Chin L, Gray JW. Translating insights from the cancer genome into clinical practice. Nature 2008;452:553–63.
[25] Rader DJ, Daugherty A. Translating molecular discoveries into new therapies for atherosclerosis. Nature 2008;451:904–13.
[26] Ormond KE, Wheeler MT, Hudgins L, Klein TE, Butte AJ, Altman RB, et al. Challenges in the clinical application of whole-genome sequencing. Lancet 2010;375:1749–51.
[27] Statnikov A, Aliferis CF. Analysis and computational dissection of molecular signature multiplicity. PLoS Comput Biol 2010;6:e1000790.
[28] Zakharkin SO, Kim K, Mehta T, Chen L, Barnes S, Scheirer KE, et al. Sources of variation in affymetrix microarray experiments. BMC Bioinform 2005;6:214.
[29] Qiu X, Brooks AI, Klebanov L, Yakovlev N. The effects of normalization on the correlation structure of microarray data. BMC Bioinform 2005;6:120.
[30] Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci USA 2006;103:5923–8.
[31] Roepman P, Kemmeren P, Wessels LF, Slootweg PJ, Holstege FC. Multiple robust signatures for detecting lymph node metastasis in head and neck cancer. Cancer Res 2006;66:2361–6.
[32] Consortium MAQC. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol 2010;28:827–38.
[33] Creighton CJ. Multiple oncogeneic pathway signatures shows coordinate expression patterns in human prostate tumors. PLoS ONE 2008;3:e1816.
[34] Maglietta R, Distaso A, Piepoli A, Palumbo O, Carella M, D'Addabbo A, et al. On the reproducibility of results of pathway analysis in genome-wide expression studies of colorectal cancers. J Biomed Inform 2010;43:397–406.
[35] Chen J, Sam L, Huang Y, Lee Y, Li J, Liu Y, et al. Protein interaction network underpins concordant prognosis among heterogeneous breast cancer signatures. J Biomed Inform 2010;43:385–96.
[36] Manoli T, Gretz N, Gröne HJ, Kenzelmann M, Eils R, Brors B. Group testing for pathway analysis improves comparability of different microarray datasets. Bioinformatics 2006;22:2500–6.
[37] Barth AS, Kuner R, Buness A, Ruschhaupt M, Merk S, Zwermann L, et al. Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies. J Am Coll Cardiol 2006;48:1610–7.
[38] Kittleson MM, Minhas KM, Irizarry RA, Ye SQ, Edness G, Breton E, et al. Gene expression analysis of ischemic and nonischemic cardiomyopathy, shared and distinct genes in the development of heart failure. Physiol Genomics 2005;21:299–307.
[39] Hannenhalli S, Putt ME, Gilmore JM, Wang J, Parmacek MS, Epstein JA, et al. Transcriptional genomics associates FOX transcription factors with human heart failure. Circulation 2006;114:1269–76.
[40] Wang H, Zheng H, Azuaje F. Ontology- and graph-based similarity assessment in biological networks. Bioinformatics 2010;26:2643–4.
[41] Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004;3.
[42] Chini A, Fonseca S, Fernández G, Adie B, Chico JM, Lorenzo O, et al. The JAZ family of repressors is the missing link in jasmonate signalling. Nature 2007;448:659–60.
[43] Ihaka R, Gentleman R. A language for data analysis and graphics. J Comput Graphical Stat 1996;5:299–314.
[44] Kaimal V, Bardes EE, Tabar SC, Jegga AG, Aronow BJ. ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. Nucleic Acids Res 2010;38:96–102.
[45] Frank E, Hall M, Trigg L, Holmes G, Witten IH. Data mining in bioinformatics using Weka. Bioinformatics 2004;20:2479–81.
[46] Witten IH, Frank E. Data mining. Practical machine learning tools and technique. San Francisco: Morgan Kaufman; 2005.
[47] Platt J. Sequential minimal optimization: a fast algorithm for training support vector machines. Microsoft research technical report, MSR-TR-98-14; 1998.
[48] Klebanov L, Yakovlev A. How high is the level of technical noise in microarray data? Biol Direct 2007;2:9.
[49] Popovici V, Chen W, Gallas BG, Hatzis C, Shi W, Samuelson FW, et al. Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. Breast Cancer Res 2010;12:R5.
[50] Yamaura G, Turoczi T, Yamamoto F, Siddqui MAQ, Maulik N, Das Dipak K. STAT signaling in ischemic heart: a role of STAT5A in ischemic preconditioning. Am J Physiol Heart Circ Physiol 2003;285:H476–82.
[51] Siow RC, Sato H, Mann GE. Heme oxygenase-carbon monoxide signalling pathway in atherosclerosis: anti-atherogenic actions of bilirubin and carbon monoxide? Cardiovasc Res 1999;41:385–94.
[52] Eriksson M, Jokinen E, Sistonen L, Leppä S. Heat shock factor 2 is activated during mouse heart development. Int J Dev Biol 2000;44:471–7.
[53] Takeda N, Manabe I, Uchino Y, Nagai R. Significance of the transcription factor Klf5 in myocardial hypertrophy in response to pressure overload. Circulation 2008;118:S_431.
[54] Leshnower BG, Sakamoto H, Zeeshan A, Parish LM, Hinmon R, Plappert T, et al. Role of acetaminophen in acute myocardial infarction. Am J Physiol Heart Circ Physiol 2006;290:H2424–31.
[55] Felkin LE, Birks EJ, George R, Wong S, Khaghani A, Yacoub MH, et al. A quantitative gene expression profile of matrix metalloproteinases (MMPS) and their inhibitors (TIMPS) in the myocardium of patients with deteriorating heart failure requiring left ventricular assist device support. J Heart Lung Transplant 2006;25:1413–9.
[56] Statnikov L, Wang CF, Aliferis A. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinform 2008;9:319.
[57] Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics 2008;9.
[58] Azuaje F, Devaux Y, Wagner DR. Coordinated modular functionality and prognostic potential of a heart failure biomarker-driven interaction network. BMC Syst Biol 2010;4:6.
[59] Azuaje F, Devaux Y, Vausort M, Yvorra C, Wagner DR. Transcriptional networks characterize ventricular dysfunction after myocardial infarction: a proof-of-concept investigation. J Biomed Inform 2010;43:812–9.
[60] Nam D, Kim SY. Gene-set approach for expression pattern analysis. Brief Bioinform 2008;9:189–97.
[61] Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005;365:488–92.